Available online at www.sciencedirect.com
ScienceDirect
Speech Communication 53 (2011) 36–50
SPEECH COMMUNICATION
www.elsevier.com/locate/specom

Constructing a spoken dialogue corpus for studying paralinguistic information in expressive conversation and analyzing its statistical/acoustic characteristics

Hiroki Mori [a,*], Tomoyuki Satake [a], Makoto Nakamura [b], Hideki Kasuya [a]

[a] Graduate School of Engineering, Utsunomiya University, 7-1-2, Yoto, Utsunomiya-shi 321-8585, Japan
[b] Faculty of International Studies, Utsunomiya University, 350, Minemachi, Utsunomiya-shi 321-8505, Japan

Received 17 November 2009; received in revised form 9 July 2010; accepted 3 August 2010

Abstract

The Utsunomiya University (UU) Spoken Dialogue Database for Paralinguistic Information Studies is introduced. The UU Database is especially intended for use in understanding the usage, structure and effect of paralinguistic information in expressive Japanese conversational speech. Paralinguistic information refers to meaningful information, such as emotion or attitude, delivered along with linguistic messages. The UU Database comes with labels of perceived emotional states for all utterances. The emotional states were annotated with six abstract dimensions: pleasant–unpleasant, aroused–sleepy, dominant–submissive, credible–doubtful, interested–indifferent, and positive–negative. To stimulate expressively-rich and vivid conversation, the "4-frame cartoon sorting task" was devised. In this task, four cards each containing one frame extracted from a cartoon are shuffled, and each participant with two cards out of the four then has to estimate the original order. The effectiveness of the method was supported by a broad distribution of subjective emotional state ratings. Preliminary annotation experiments by a large number of annotators confirmed that most annotators could provide fairly consistent ratings for a repeated identical stimulus, and the inter-rater agreement was good ($W \simeq 0.5$) for three of the six dimensions. Based on the results, three annotators were selected for labeling all 4840 utterances. The high degree of agreement was verified using such measures as Kendall's W. The results of correlation analyses showed that not only prosodic parameters such as intensity and f_0 but also a voice quality parameter were related to the dimensions. Multiple correlation of above 0.7 and RMS error of about 0.6 were obtained for the recognition of some dimensions using linear combinations of the speech parameters. Overall, the perceived emotional states of speakers can be accurately estimated from the speech parameters in most cases.

Keywords: Emotional state; Expressive speech; Annotation; Abstract dimensions; Spontaneous speech; Spoken dialogue

1. Introduction

Paralinguistic information generally refers to information that is not linguistic content itself but some meaningful information delivered along with linguistic messages. Sometimes paralinguistic information is even more eloquent than the linguistic message itself. Research on paralinguistic information is attracting growing attention among the speech science community, partly because revealing the nature of paralinguistic information involves not only advanced speech technologies such as human-like agents, but also revealing the hidden nature of speech communication between humans.

Emotion and expressivity in speech is a central topic of paralinguistic information (Cowie et al., 2001; Erickson, 2005). However, paralinguistic information should not be viewed only within the context of biological effects of individuals as in the traditional psychology of emotion, because

* Corresponding author. Tel.: +81 28 689 6120; fax: +81 28 689 6119.
E-mail address: hiroki@klab.ee.utsunomiya-u.ac.jp (H. Mori).

0167-6393/$ - see front matter © 2010 Elsevier B.V. All rights reserved.
doi:10.1016/j.specom.2010.08.002

speech has an important function: interaction. If the aim of a study on paralinguistic information includes not only its emotional aspects but also its socio-linguistic roles, scripted corpus collection is almost useless. Today, there is a trend toward considering natural emotion (Douglas-Cowie et al., 2003; Aubergé et al., 2003) in corpus development for expressive speech studies. Recent works on developing a spontaneous speech corpus for paralinguistic information studies include Greasley et al. (1995), Douglas-Cowie et al. (2000), Campbell (2003), Devillers and Vidrascu (2006), Truong et al. (2008), and Arimoto et al. (2008).

However, there is a serious problem in all attempts to design a speech corpus with natural emotion, i.e. how to label emotions in the corpus? Unlike traditional emotion studies, well-established emotion categories do not apply to most utterances in daily conversations. An alternative scheme is therefore needed for describing the emotional states that are expressed in speech (Cowie and Cornelius, 2003). Possible approaches include lists of key emotions, dimensional representations of underlying emotion, and physiological measures such as heart rate, eye blink, EEG, or facial muscle activities. Among these, effect-oriented emotion labels, which describe the listener's impression, are indispensable if the corpus is to be used for building speech applications such as expressive speech synthesis, or exploring the interactive effects of paralinguistic information in speech communication. Irrespective of the application, establishing a common ground for labeling speech expressivity is certainly a major challenge. One of the objectives of our corpus is to provide a reference implementation for future corpus development.

The Utsunomiya University Spoken Dialogue Database for Paralinguistic Information Studies (UU Database) is the first public speech corpus specifically designed for studies on paralinguistic information in expressive Japanese dialogue speech. The UU Database is characterized by its unique task design and emotional state labels. Other existing corpora of natural dialogue speech (e.g. Reading/Leeds Emotion in Speech Corpus (Greasley et al., 1995), the Belfast Naturalistic Emotional Database (Douglas-Cowie et al., 2000), and the JST/CREST ESP Corpus (Campbell, 2003)) are not available for public use. Dialogues provided by some Japanese speech corpora (e.g. PASD, the Japanese Map Task Corpus (Horiuchi et al., 1999), and ATR-SLDB (Morimoto et al., 1994)) are too restrictive, or businesslike, for studies of speech expressivity; in fact, there is no public Japanese spoken dialogue corpus that provides speech with natural emotion. The list of relevant corpora is shown in Table 1. Despite the variation in terminology (evaluation = valence = pleasantness, activation = arousal), the table shows that the dimensional description has been adopted as a de facto standard for annotating natural emotion in recent studies. Only the Vera am Mittag Database and the UU Database are public corpora provided with dimension-based emotion annotation. Because most emotions appearing in the Vera am Mittag Database were

Table 1
Existing speech corpora for studying spontaneous dialogue.

Corpus	Language	Data collection method	Model of emotion annotation	Public availability
HCRC map task corpus (Anderson et al., 1991)	English	Map task	Nothing	Available
Japanese map task corpus (Horiuchi et al., 1999)	Japanese	Map task	Nothing	Freely available
PASD simulated spoken dialogue corpus	Japanese	Simulated dialogue under various tasks	Nothing	Freely available
ATR-SLDB (Morimoto et al., 1994)	Japanese	Simulated dialogue under the hotel reservation task	Nothing	Available
CallHome (LDC)	English, Arabic, German, Spanish, Japanese, Chinese	Telephone call to family members or close friends	Nothing	Available
Reading/leeds emotion in speech corpus (Greasley et al., 1995)	English	Interviews on radio/TV programs	Category (happiness, sadness, anger, fear, disgust), freely choiced label	Not available
Belfast naturalistic database (Douglas-Cowie et al., 2000)	English	Interviews on TV chat shows	Dimension (activation–evaluation, rated using Feeltrace), Category (40 words)	Not available
JST/CREST ESP corpus (Campbell, 2003)	Japanese	Daily recording of volunteers' natural spoken interactions	Dimension (activation-evaluation, rated using Feeltrace)	Not available
The Vera am Mittag German audio–visual spontaneous speech database (Grimm et al., 2008)	German	Spontaneous and emotional speech recorded from a TV talk show	Dimension (valence, activation, and dominance, using the self assessment manikins)	Freely available
TNO-Gaming corpus (Truong et al., 2008)	Dutch	Talks with friends during playing multiplayer video game with preset events	Dimension (arousal, valence), Category (12 words) Self and observer ratings	Not available
UU database (this paper)	Japanese	Four-frame cartoon sorting task	Dimension (pleasantness, arousal, dominance, credibility, interest, and positivity)	Freely available